DOCUMENTATION
INCORPORATED

SPECIALISTS IN BASIC AND APPLIED INFORMATION THEORY

2521 CONNECTICUT AVENUE, N. W.
WASHINGTON 8, D. C.
COLUMBIA 5-4577

COMMUNICATION THEORY AND STORAGE AND RETRIEVAL SYSTEMS

TECHNICAL REPORT NO. 12

Prepared under
Contract Nonr-1305(00)

FOR THE OFFICE OF NAVAL RESEARCH

October
1955

RECORDS ENGINEERING MEMORY DEVICES INFORMATION SYSTEMS
Developers of: The UNITERM System of Coordinate Indexing - RADEX Random Filing Cards - MATREX Indexing Machines
MEMATREX Association of Ideas Machines

DOCUMENTATION
INCORPORATED

SPECIALISTS IN BASIC AND APPLIED INFORMATION THEORY

2521 CONNECTICUT AVENUE, N. W.
WASHINGTON 8, D. C.
COLUMBIA 5-4577

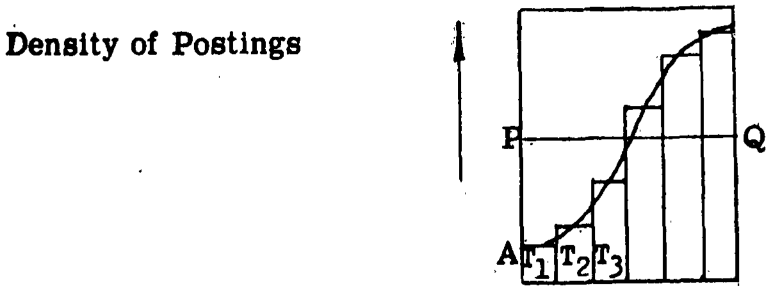
COMMUNICATION THEORY AND STORAGE AND RETRIEVAL SYSTEMS

TECHNICAL REPORT NO. 12

The research we are doing for the Office of Naval Research in the theory of those special types of information systems whose primary purpose is to store and retrieve information has led, on one hand, to the design of various pieces of equipment and, on the other, to the development of new approaches to the fundamental theory of storage and retrieval systems. In this article, we will indicate in general terms certain relations of this embryonic theory of storage and retrieval systems to the mathematical theory of communication developed by Claude Shannon and others. We say embryonic here because whatever theory of storage and retrieval systems exists is almost wholly qualitative and has not yet found its Shannon who might give it formal expression. Most of the concepts which are basic in communication theory are also basic in storage and retrieval theory, e.g., information, entropy, redundancy, noise, capacity, probability, etc., but we do not yet know the mathematical relations of these concepts in storage and retrieval theory. We would expect a generalized information theory to cover both communication theory and storage and retrieval theory as special aspects -- just as a generalized wave theory might cover both electromagnetism and light. But until such a generalized theory is developed, we can find in the more highly developed special field certain suggestions concerning problems in the less developed field.

A collation system approaches maximum efficiency when the two series which are being collated approach equality, and items in one series are randomly distributed through the other. Similarly, any system of storing data is utilized to maximum efficiency when all storage elements are equally loaded. A library classification system ideally should have an equal number of items in each class and subclass. An IBM card should have all its fields used in relatively equal proportions. The new Kodak Minicard System which has storage elements for pieces of film is only efficient when the number of pieces of film in each element is relatively equal to all the other elements. In our own work with systems of coordinate indexing, e. g., the Uniterm System, the Batten System, and the Matrex System, we have utilized several devices to increase the density of posting on lightly posted cards and cut the density on heavily posted cards, i. e., to equalize the storage elements.

An ideal system would exhibit a curve equivalent to the horizontal line PQ in the following figure; in actual fact we always find a curve which resembles AB:



In communication theory, we find an interesting parallel:

> If one is concerned, as in a simple case, with a set of n independent symbols, or a set of n independent complete messages for that matter, whose probabilities of choice are p_1, p_2 . . . p_n, then the actual expression for the information is
>
> $$H = -[p_1 \overset{*}{\log} p_1 + p_2 \log p_2 + \ . \ . \ . + p_n \log p_n],$$
>
> or
>
> $$H = -\Sigma p_i \log p_i$$
>
> *log to the base 2
>
> Where the symbol Σ indicates, as is usual in mathematics, that one is to sum all terms like the typical one, $p_i \log p_i$, written as a defining sample . . . . This looks a little complicated; but let us see how this expression behaves in some simple cases. . . . Suppose first that we are choosing only between two possible messages, whose probabilities are then p_1 for the first and $p_2 = 1 - p_1$ for the other. If one reckons, for this case, the numerical value of H, it turns out that H has its largest value, namely one, when the two messages are equally probable; that is to say when $p_1 = p_2 = \frac{1}{2}$; that is to say, when one is completely free to choose between the two messages. Just as soon as one message becomes more probable than the other (p_1 greater than p_2, say), the value of H decreases. And when one message is very probable (p_1 almost one and p_2 almost zero, say), the value of H is very small (almost zero).[1]

If, in a storage and retrieval system, we let S equal the measure of efficient storage or loading and P equal, not the probability of a message,

[1] Shannon, Claude and Weaver, Warren, The Mathematical Theory of Communication (The University of Illinois Press, Urbana, 1949), p. 105.

but the amount of loading at each position in the system, we can write

$$S = -\Sigma\, p_i \log p_i$$

for a given n (number of positions) S is a maximum and equal to log n when all the p_i are equal, i.e., $\frac{1}{n}$.[2] Thus, the formula for efficiency of storage is exactly analogous to the formula for the amount of information.

If the measure of the information in a system of messages is H, the redundancy in the system is 1 - H. It follows that if the amount of information is maximized when all messages are equally probable, then the redundancy in the system is maximized when there is the largest variation in probability of the messages, e.g., one message has the probability 1 and all others have the probability 0. According to Shannon, ordinary English text has a redundancy of approximately 50%, the extent to which the succession of letters in any English message departs from random distribution.

In sending a message, sending a "u" after "q" is completely redundant; it gives no information to the recipient that he didn't have when he received "q". We must be careful to avoid equating redundancy with useless information. In perfect communication systems, this equivalence would hold, but the existence of "noise" in communication systems makes some degree of redundancy essential. The "u" following "q" is redundant, but suppose we eliminated it in our messages. Then the word "queer" would be sent as "qeer". But nowsuppose that noise in the

[2] *Ibid.*, p. 21

the communication channel obscured the transmission of the "q". If the recipient has only "eer", he is obviously in a much worse position than if he had "ueer".

It remains true that where proper coding transmission and other devices can reduce noise, it becomes the aim of designers of communication, instrumentation, and storage and retrieval systems to eliminate redundancy and thereby to increase the information capability of the system.

If we store items of information under the following items:

dog	brown	houses
A	B	C

there is less redundancy in our system than if we had used

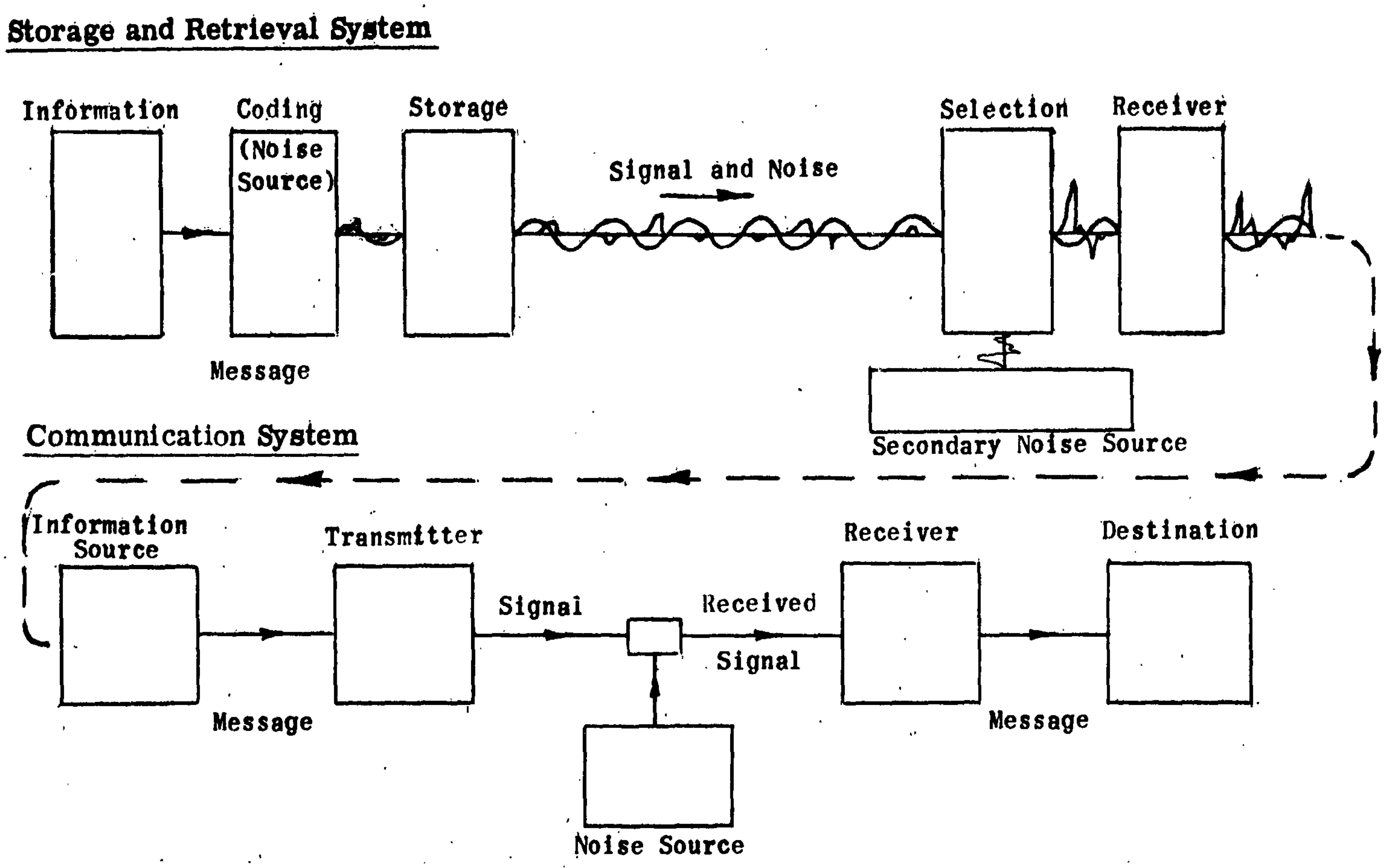
dog	brown	houses	brownhouses	housedogs	dog houses	brown dogs
A	B	C	D	E	F	G

On the other hand, a system with only the terms A, B, C will deliver some "noise" in its retrieval process. One of the best ways to sum up the nature of coordinate indexing is to note that such systems eliminate the redundancy of word order and word combinations in storing information at the price of a certain percentage of noise in the retrieval process. Similarly, when we go from the words of a Uniterm or Batten System to letter elements in the Matrex System, we carry the elimination of redundancy one step further and thereby increase the percentage of noise in the retrieval process.

At this point we must note again the absence of a generalized information theory and pause in our search for analogues between the special theory of communication and the theory of storage and retrieval systems. Because here with this matter of noise, the analogy breaks down. Noise in a communication system does not arise from the elimination of redundancy but from the limitation of channel capacity or external conditions which effect a communication system (e. g., background radiation, electric storms, etc.) But in a storage and retrieval system, we introduce noise when we index or code information for economical storage.

This notion of channel capacity again has no very clear analogue in storage and retrieval systems. Storage capacity is not analogous to the channel of a communication system but to the transmitter in which the information or message is converted to a signal. We can show this by presenting the outline of a communication system as given by Shannon and Weaver[3] and a similar outline of a storage and retrieval system. Information selected from a storage and retrieval system may constitute the input of a communication system.

[3] Ibid., p. 98



In communication systems the message is coded (made compatible with the channel) at the transmitter. In storage and retrieval systems, information is coded for storage. In more philosophic terms, we would say experience is given a symbolic expression and the failure -- the necessary failure of any and every symbolic system to capture living reality -- introduces noise into the information or communication process. Actually, the "noise" introduced by a special device or system is only a special case of "noise" in the general or all pervasive sense.

If we consider isolated systems at certain levels of abstraction, we can consider "noise" an accident of speech, coding, or channel characteristics,etc., and we can evaluate devices in terms of the amount of noise they introduce or eliminate. Shannon makes such an abstraction and concerns himself with an isolated system when he says: "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem."[4]

The semantic problem, the problem of pervasive noise in all systems of symbolic communication, is something to be understood philosophically and is irrelevant to the engineering aspects which concern the particular noise introduced by special devices. Mr. Weaver, who sees in the mathematical theory of communication a powerful analytical tool, believes that it does constitute a start at least towards the understanding if not the solution of the semantic problem. Thus he remarks cryptically that although the semantic aspects of communication are irrelevant to the engineering aspects, "this does not mean that the engineering aspects

[4] Ibid., p. 3

are necessarily irrelevant to the semantic aspects."[5] What Mr. Weaver desires to emphasize, and I think quite rightly, is that from a special concern with designing the best communication channels we can gain an insight into the basic semantic problems of the meaning and effectiveness of communication. For example, we are never so much aware of the ambiguity of ordinary speech as we are when we are concerned with devising a special language. When we introduce noise in a storage system by superimposed coding, we can recognize what we have done as a special case of the kind of semantic confusion we get when someone is "bursting with ideas and can't find enough words to express himself." Actually, stuttering may be a primitive form of noise arising out of superimposition.

The fact that we come to ultimate semantic or philosophic issues when considering storage and retrieval systems rather than communication systems indicates that the former is more fundamental -- less abstract -- than the latter. This conclusion is strengthened by noting that we have a mathematical theory of communication while we still await a mathematical theory of storage and retrieval. Perhaps Shannon himself will go on to give us this more general theory.

This conclusion is strengthened even more by the implications of another observation made by Mr. Weaver: "The information source selects a desired message out of a set of possible messages. The selected

[5] *Ibid.*, pp. 99-100

message may consist of written or spoken words, or of pictures, music, etc."[6] This "selection" must be a selection from storage, if we understand storage to encompass a living memory, a vocabulary, a language, an organized library, the patent office, a code book or even the collection of greetings maintained at telegraph offices. In all of these things and in many more like them we store symbolically, which is to say more or less adequately, the meanings we experience and live by.

Respectfully submitted,

Mortimer Taube

Mortimer Taube
President

[6] *Ibid.*, p. 98